Bachelor of Science in Computer Science & Engineering

Multiple Disease Prediction System Using Machine Learning Model

by

Salman Farsi

ID: 1804102

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

07 February, 2023

Chittagong University of Engineering & Technology (CUET)

Department of Computer Science & Engineering

Chattogram-4349, Bangladesh.

Final project report

on

Multiple Disease Prediction System Using Machine Learning Model

Student Name	Salman Farsi	Session : 2021-2022
ID	1804102	

Supervisor Name	: Ashim Dey
Designation	: Assistant Professor Department of Computer Science & Engineering

Supervisor Name	: Avishek Das
Designation	: Lecturer Department of Computer Science & Engineering

Department	: Computer Science & Engineering
Program	: B.Sc. Engineering

Title	**: Multiple Disease Prediction System Using Machine Learning Model**

Table of Contents

List of Figures

List of Tables

1 Introduction

There are numerous machine learning approaches that can perform predictive analytic on vast volumes of data in a range of businesses. Although using predictive analytic in healthcare is more challenging. Globally, Diseases like diabetes, heart-related diseases, and parkinson's disease are causing many deaths but most of these deaths are due to the lack of timely check-ups of the diseases. As a result, many lives can be saved by early detection and diagnosis of these disorders. According to the International Diabetes Federation, there are 285 million diabetic people worldwide.The results obtained using machine learning classifiers like **Support Vector Machine, Random Forest, Logistic Regression** can be successfully used for diagnosing diseases with higher accuracy like, diabetes, heart, parkinson and several other diseases. Though sometimes, the doctor itself can diagnose. But when it comes to analyzing the relationship between several diagnostic test results, there is no alternative to data science. Besides the early prediction can aware the people about their health condition. For example, healthcare professionals working in the area of cardiac disease have their own limits and can not forecast the probability of high accuracy in cardiac diseases as it may happen abruptly to any person. Looking at different diagnostic results the machine learning model can be fed by this value to create the prediction. Because for each disease there exist several disease that has some common relationship with another particular diagnostic value. The main goal of this project is to use the above mentioned machine learning classification model to anticipate these dangerous diseases and to make an online based application that uses the idea of machine learning to make predictions which will be accessible to the general public. Also to analyze the accuracy using K-Fold Cross Validation.

2 Outline of Methodology

In my proposed work, I aim to predict multiple diseases in one system.The below workflow which will be followed to predict the diseases.

- At first, some data related to the particular disease need to be collected. I will try to train my model using the labelled data set. So, I must **preprocess the data** before attempting to evaluate it. And analyze the value of different features like **mean, median, standard deviation, percentile etc.**.
- I therefore **visualized the data sets** through several plotting technique like scatter plot, bar chart and histogram. Here data visualization is important for finding the co-relation between the different diagnostic test values.
- Then for **supervised learning** purpose I separated the features and the target column from the dataset.
- After that I must **standardize the data** set using **Standard Scalar** because the values in several features are very much skewed and need to be adjusted in a common range to get the better accuracy from the model.
- Here train test and split method was used to divide the dataset into trained data(80%) and test data(20%).
- Then the **K-fold cross validation** is being used to find out how well the machine learning model can predict the outcome of the unseen data and the best classifier will be chosen. In the k-fold cross validation we used three different classifier namely **Support Vector Machine(SVM), Random Forest and Logistic Regreession** to train our model. There I choose the best classifer.
- Finally, our machine learning model will be trained that can predict whether a person has a specific disease or not when given new datas from the user.

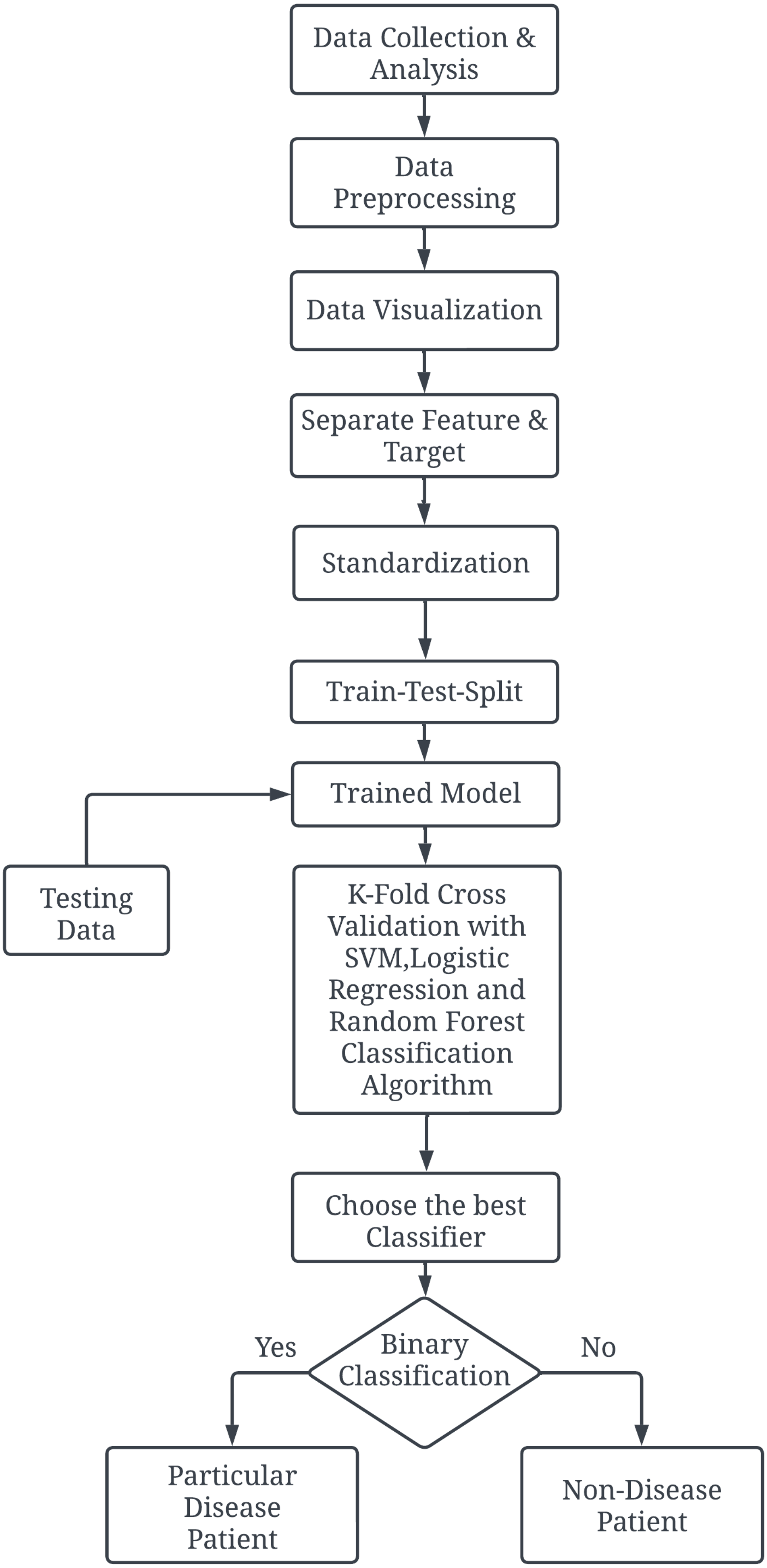

Data Collection & Analysis
Data Preprocessing
Data Visualization
Separate Feature & Target
Standardization
Train-Test-Split
Trained Model
Testing Data
K-Fold Cross Validation with SVM,Logistic Regression and Random Forest Classification Algorithm
Choose the best Classifier
Binary Classification
Yes
No
Particular Disease Patient
Non-Disease Patient

Figure 2.1: A block diagram of proposed methodology

3 Required Resources

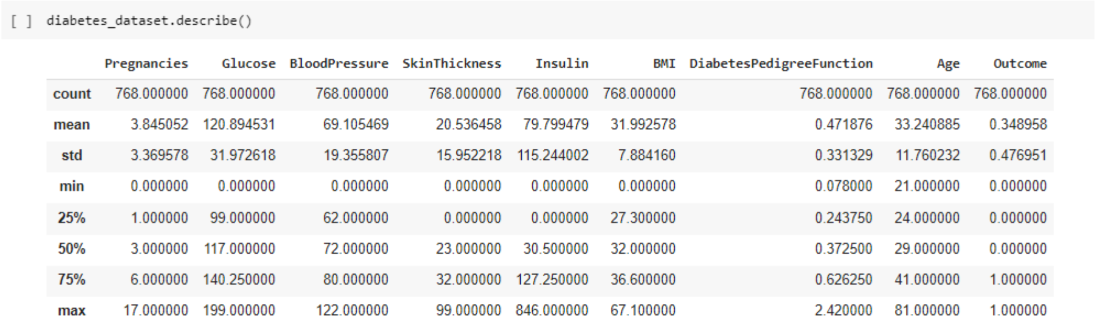
- Python
- Pandas
- Scikit-Learn
- Numpy
- Matplotlib
- seaborn
- **Google Colaboratory**
- **Streamlit for Model Deployment**

4 Results and Application Interface

4.1 Diabetes Disease:

4.1.1 Data set Analysis and Visualization:

The diabetes data set contains in total 768 rows and 9 columns. Out of these 9 columns, one column is target and rest of other columns are the feature. Here **all the features are labelled**. So, there was no need of labelling the data set separately.

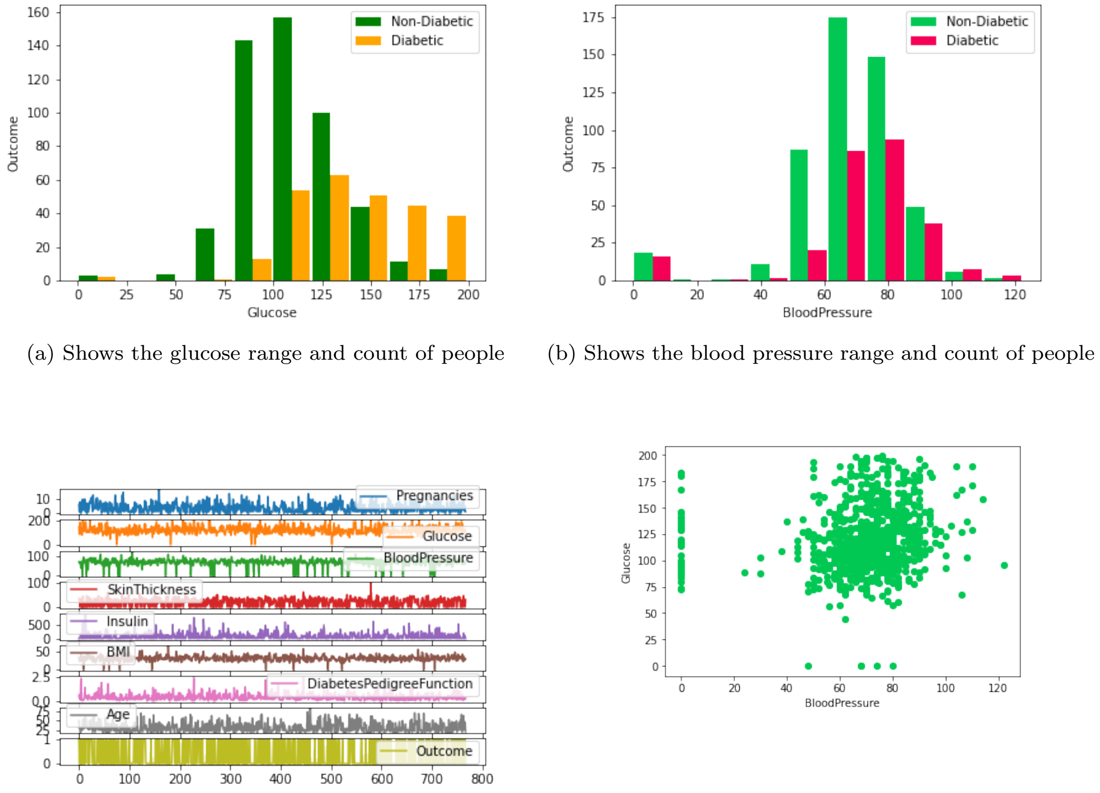


```
[ ]  diabetes_dataset.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 4.2: Details of diabetes data set

There are given some visualization of datasets that also **indicates the outlier data and regular data in the data set.**



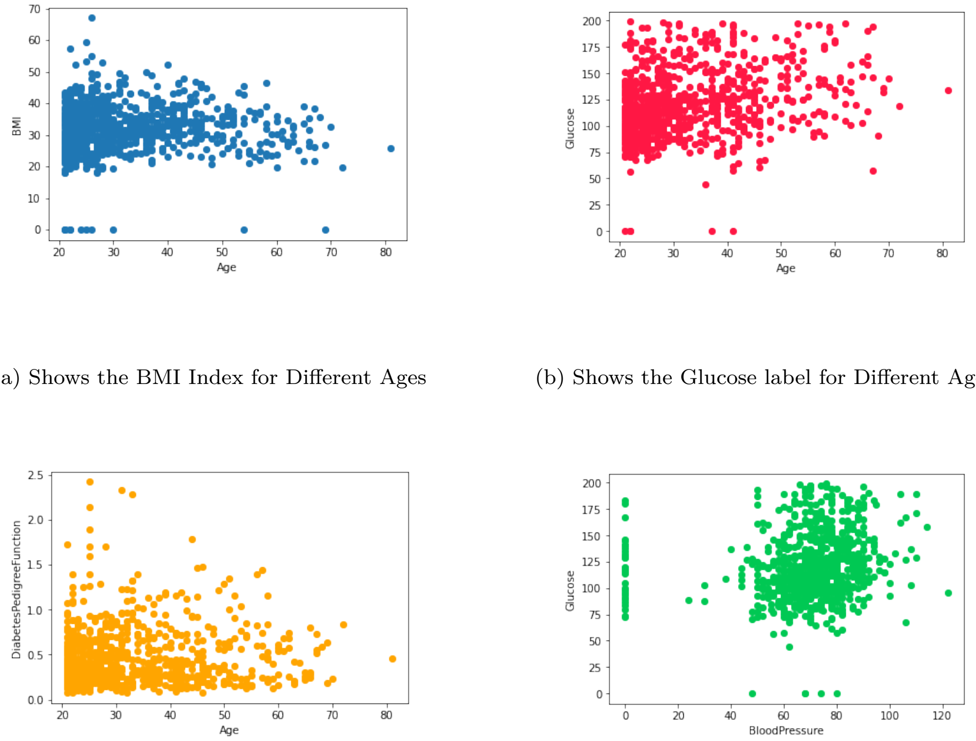
(a) Shows the glucose range and count of people

(b) Shows the blood pressure range and count of people

(c) Shows the overall range of all the features

(d) Shows the relation between blood pressure and glucose

Figure 4.3: Diabetes Data Sets Visualization



(a) Shows the BMI Index for Different Ages

(b) Shows the Glucose label for Different Ages

(c) Shows the DiabetesPedigreeFunction value for Different Ages

(d) Shows the relation between blood pressure and glucose

Figure 4.4: Visualization of diabetes data set

4.1.2 K-Fold Cross Validation Result :

Table 4.1: Accuracy Results using Different ML Models

Number of Features	8
Logistic Regression Accuracy	77.21%
Support Vector Machine Accuracy	77.08%
Random Forest Accuracy	74.73%

4.1.3 Confusion Matrix :

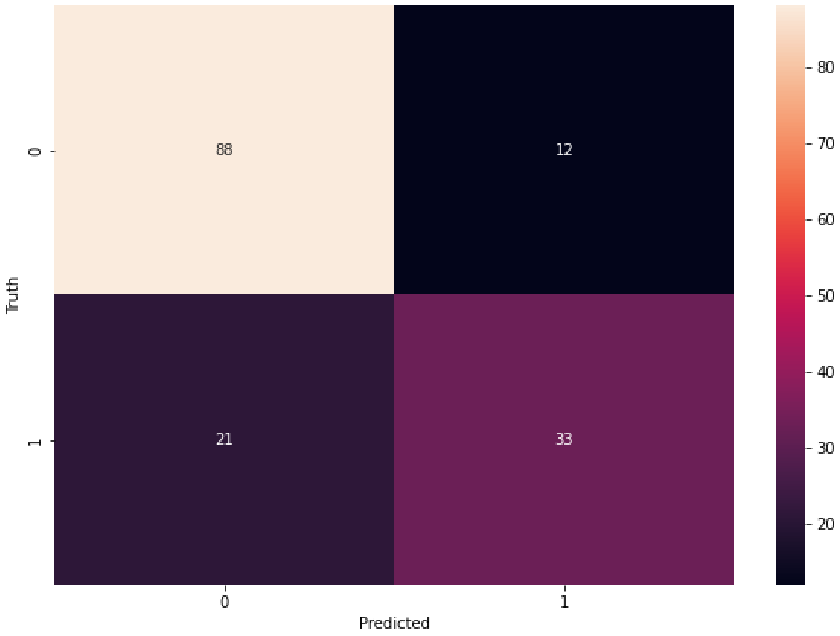


Figure 4.5: Confusion Matrix showing the contradiction between actual result and predicted result

4.1.4 Co-relation Matrix :

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.To summarize a large amount of data where the goal is to see patterns. In diabetes data set, the observable pattern is that all the variables highly correlate with each other.

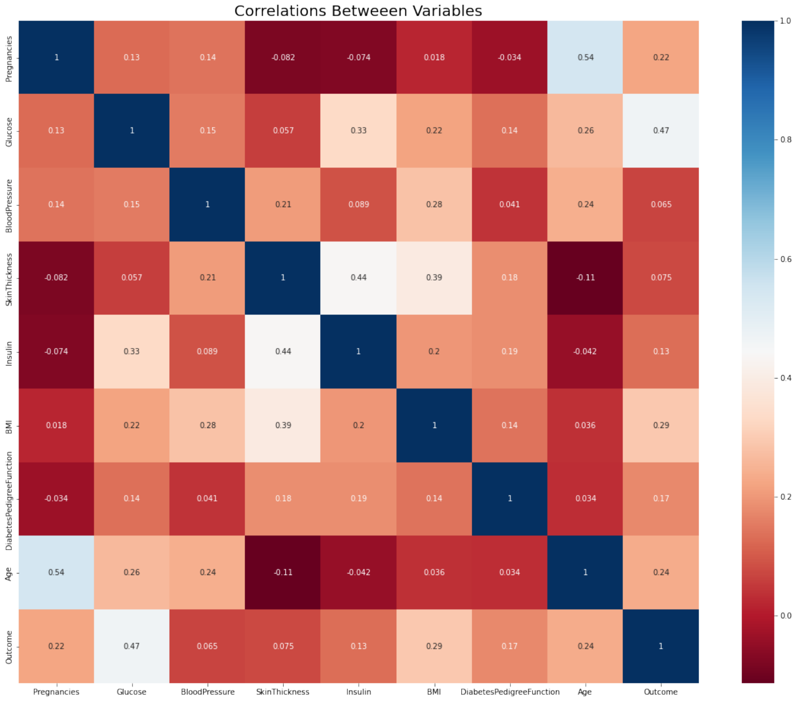
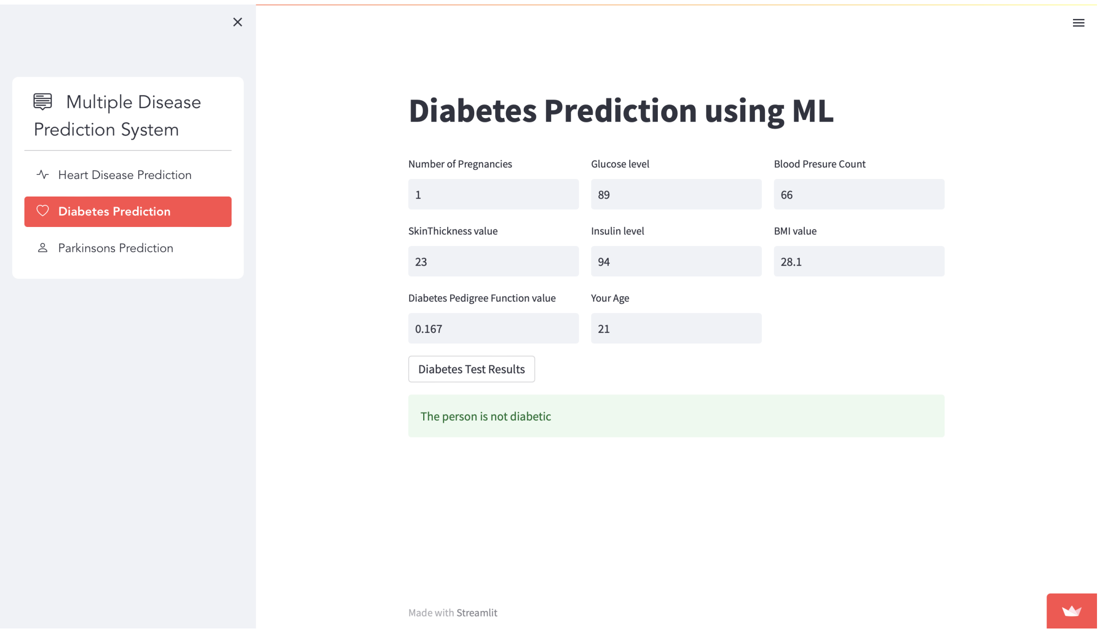


Figure 4.6: Shows the relation between all the features and their co-relationship

4.1.5 User Interface :

The patient will enter the necessary information, including the number of pregnancies, glucose level, blood pressure count, skin thickness value, insulin level, BMI value, and diabetes pedigree function value, in the diabetes prediction section. And using the trained model, our predictor model will determine whether the patient has diabetes or not.

Here, the input values for a patient who is not diabetic are shown in the text field sections of picture. The next figure also shows a picture of our predictor model interface, where the identical input data from patients are allocated, and the model predicts the outcome. And the prediction was correct.



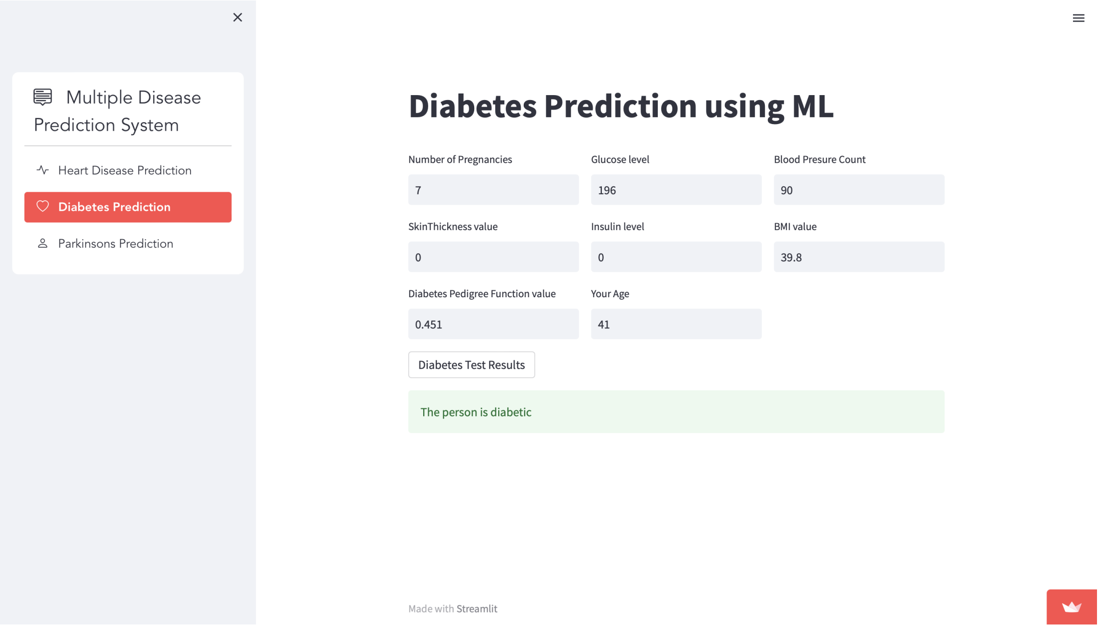
Diabetes Prediction using ML

Number of Pregnancies: 1
Glucose level: 89
Blood Presure Count: 66
SkinThickness value: 23
Insulin level: 94
BMI value: 28.1
Diabetes Pedigree Function value: 0.167
Your Age: 21

The person is not diabetic

Figure 4.7: User interface showing the prediction of non-diabetic

The below figure shows a picture of our predictor model interface, where the identical input data from patients are allocated, and the model predicts the outcome. And the forecast was correct.



Diabetes Prediction using ML

Number of Pregnancies: 7
Glucose level: 196
Blood Presure Count: 90
SkinThickness value: 0
Insulin level: 0
BMI value: 39.8
Diabetes Pedigree Function value: 0.451
Your Age: 41

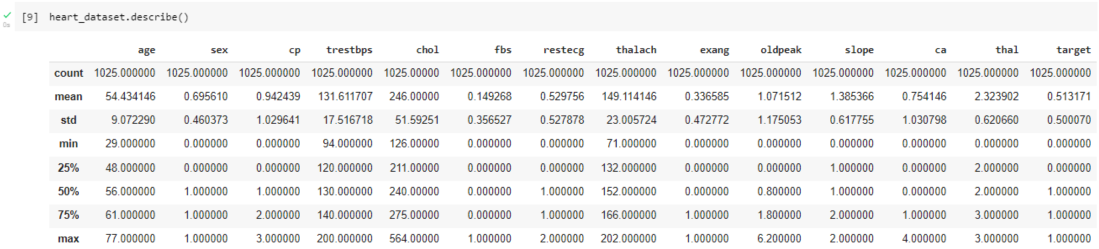
The person is diabetic

Figure 4.8: User interface showing the prediction of diabetic

4.2 Heart Disease:

4.2.1 Data Set Analysis and Visualization:

The heart disease data set contains in total 1025 rows and 14 columns. Out of these 14 columns, one column is the target and rest of other columns are the feature. Here **all the features are labelled**. So, there was no need of labelling the data set separately.

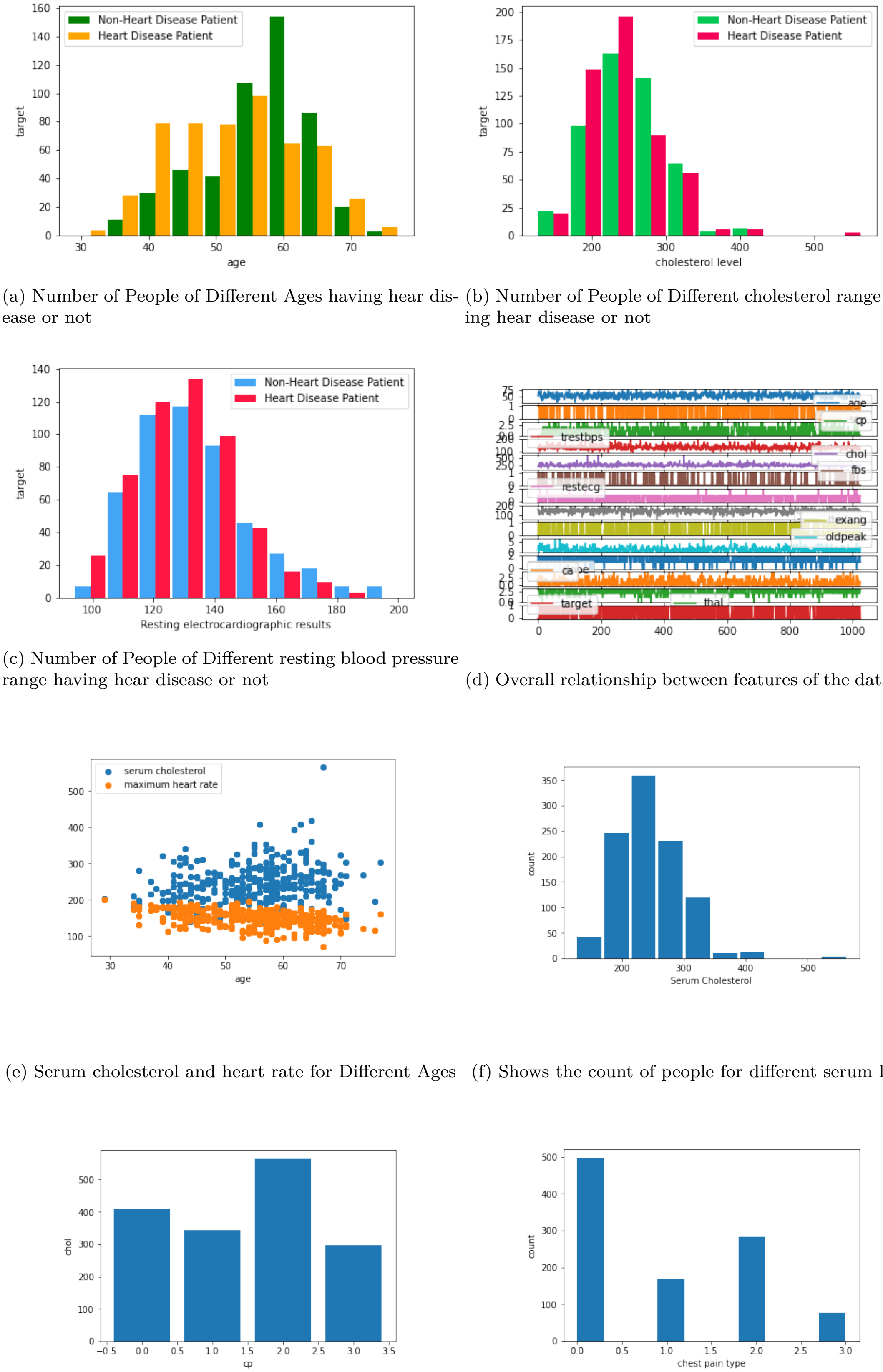


```
[9] heart_dataset.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 4.9: Details of heart disease data set

There are given some visualization of datasets that **shows relationship between different disease parameter and histogram for visualizing count of people.** This plotting shows why some types of person who are having a certain disease test result are more or less vulnerable to heart disease. It also indicates his or her future possibility of being heart disease patient. I used these plotting also to understand the classifier activities also. I used python **matplotlib** and **seaborn** library for these plotting purpose.



(a) Number of People of Different Ages having hear disease or not

(b) Number of People of Different cholesterol range having hear disease or not

(c) Number of People of Different resting blood pressure range having hear disease or not

(d) Overall relationship between features of the datasets

(e) Serum cholesterol and heart rate for Different Ages

(f) Shows the count of people for different serum level

(g) Shows the relationship between chest pain and cholesterol

(h) Shows the count of people for different chest pain level

Figure 4.10: Visualization of heart data set

4.2.2 K-Fold Cross Validation Results:

Table 4.2: Accuracy Results using Different ML Models

Number of Features	13
Logistic Regression Accuracy	84.19%
Support Vector Machine Accuracy	90.14%
Random Forest Accuracy	98.04%

4.2.3 Confusion Matrix:

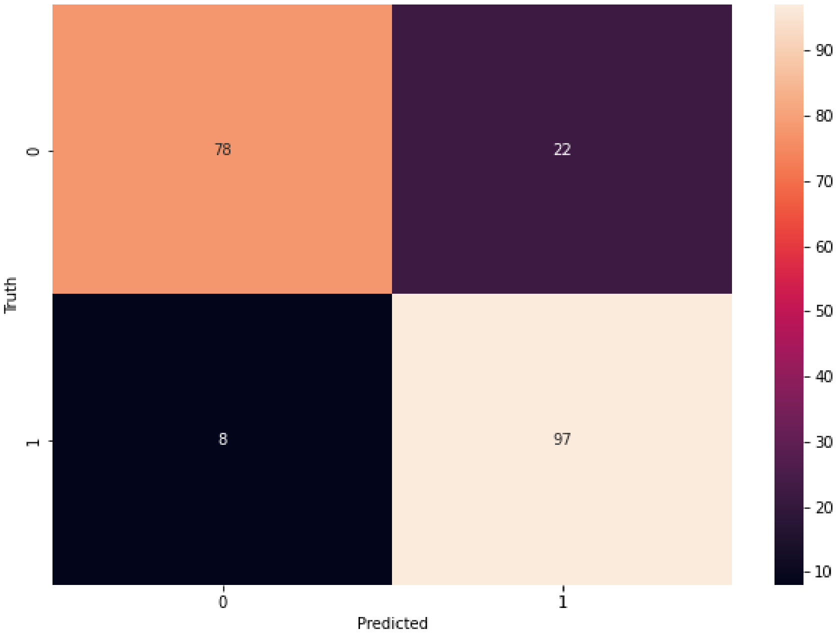


Figure 4.11: Confusion Matrix showing the contradiction between actual result and predicted result

4.2.4 Co-relation Matrix:

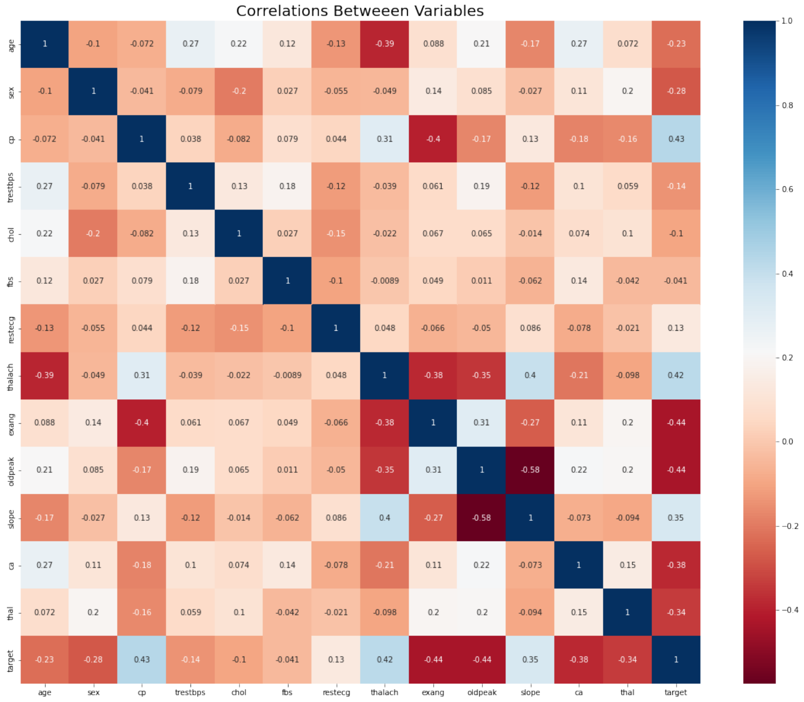
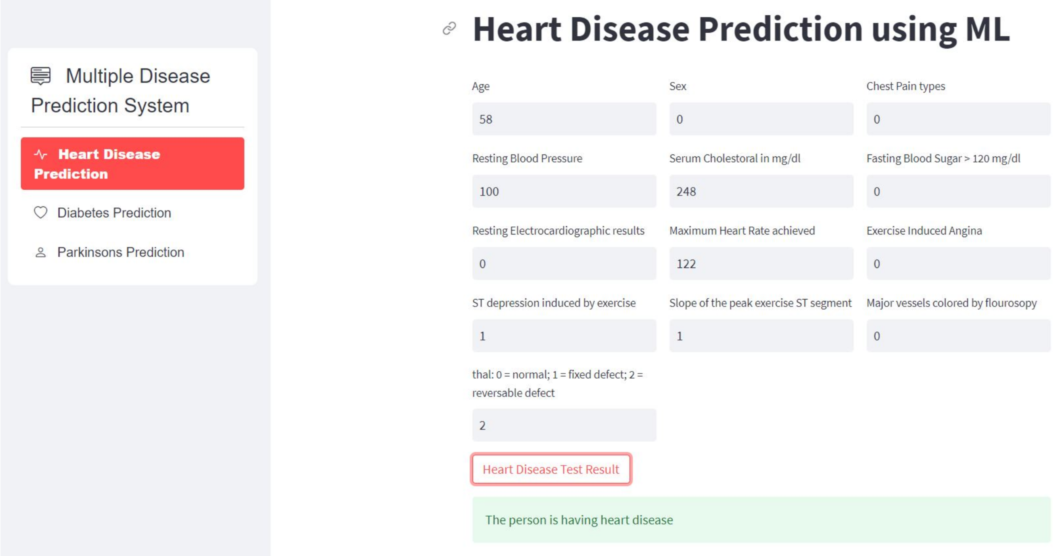


Figure 4.12: Shows the relation between all the features and their co-relationship

4.2.5 User Interface:

Heart Disease model has a total of 13 features. For the purpose of training our model, we are taking age, sex, resting blood pressure etc as features in order to predict from the data from test set if there is presence of any heart disease in our target patient. The below figure displays the presence of heart disease in a particular patient taking all necessary parameters from the patient. The figure below displays that the patient possess any form of heart disease through carefully going through the parameters obtained from the patient.



Heart Disease Prediction using ML

Age
58

Sex
0

Chest Pain types
0

Resting Blood Pressure
100

Serum Cholestoral in mg/dl
248

Fasting Blood Sugar > 120 mg/dl
0

Resting Electrocardiographic results
0

Maximum Heart Rate achieved
122

Exercise Induced Angina
0

ST depression induced by exercise
1

Slope of the peak exercise ST segment
1

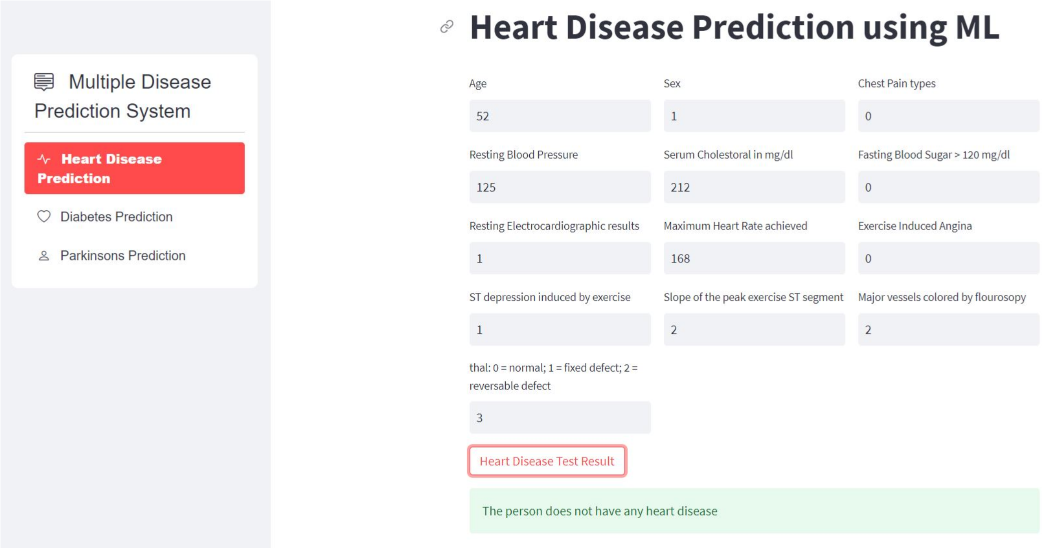
Major vessels colored by flourosopy
0

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
2

The person is having heart disease

Figure 4.13: User interface showing the prediction of heart disease patient

The figure below displays that the patient does not possess any form of heart disease through carefully going through the parameters obtained from the patient.



Heart Disease Prediction using ML

Age
52

Sex
1

Chest Pain types
0

Resting Blood Pressure
125

Serum Cholestoral in mg/dl
212

Fasting Blood Sugar > 120 mg/dl
0

Resting Electrocardiographic results
1

Maximum Heart Rate achieved
168

Exercise Induced Angina
0

ST depression induced by exercise
1

Slope of the peak exercise ST segment
2

Major vessels colored by flourosopy
2

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
3

The person does not have any heart disease

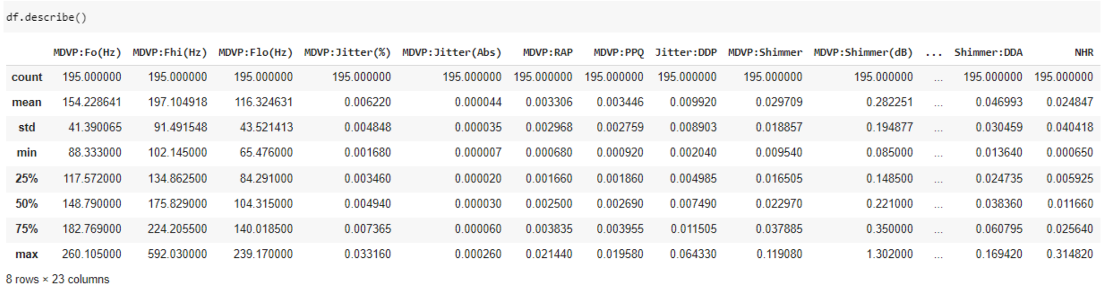
Figure 4.14: User interface showing the prediction of non heart disease patient

4.3 Parkinson Disease:

4.3.1 Data set Analysis and Visualization:

The parkinson data set contains in total 195 rows and 24 columns. Out of these 24 columns, one column is target, another one is the patient id and rest of other columns are the feature. Here **all the features are labelled**. So, there was no need of labelling the data set separately.

Parkinson disease model has total 22 features namely, Average vocal fundamental frequency, Maximum vocal fundamental frequency, Minimum vocal fundamental frequency, Several measures of variation in fundamental frequency, Several measures of variation in fundamental frequency, Several measures of variation in fundamental frequency, Several measures of variation in fundamental frequency, Several measures of variation in fundamental frequency etc.



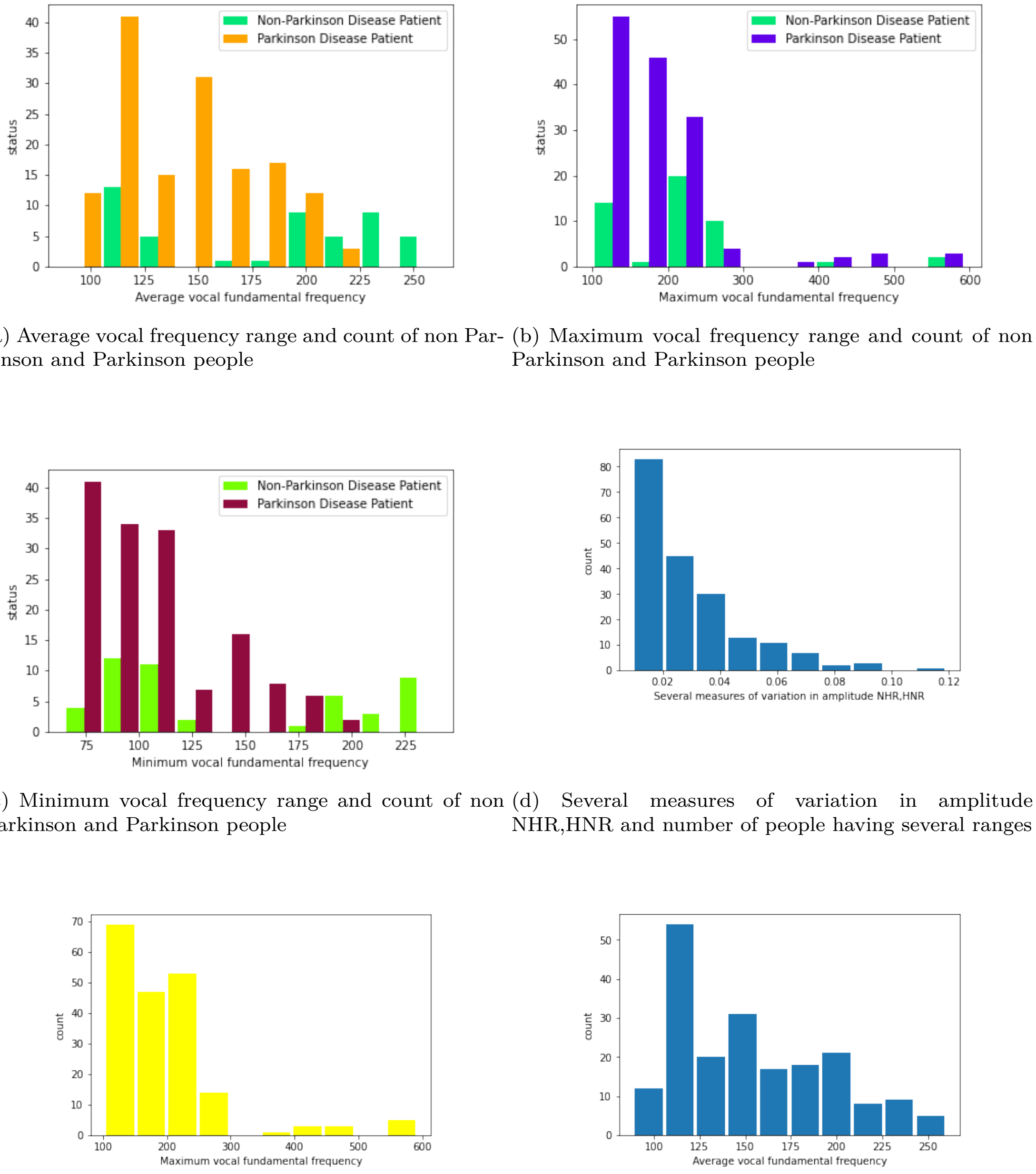
```
df.describe()
```

	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB)	...	Shimmer:DDA	NHR
count	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	...	195.000000	195.000000
mean	154.228641	197.104918	116.324631	0.006220	0.000044	0.003306	0.003446	0.009920	0.029709	0.282251	...	0.046993	0.024847
std	41.390065	91.491548	43.521413	0.004848	0.000035	0.002968	0.002759	0.008903	0.018857	0.194877	...	0.030459	0.040418
min	88.333000	102.145000	65.476000	0.001680	0.000007	0.000680	0.000920	0.002040	0.009540	0.085000	...	0.013640	0.000650
25%	117.572000	134.862500	84.291000	0.003460	0.000020	0.001660	0.001860	0.004985	0.016505	0.148500	...	0.024735	0.005925
50%	148.790000	175.829000	104.315000	0.004940	0.000030	0.002500	0.002690	0.007490	0.022970	0.221000	...	0.038360	0.011660
75%	182.769000	224.205500	140.018500	0.007365	0.000060	0.003835	0.003955	0.011505	0.037885	0.350000	...	0.060795	0.025640
max	260.105000	592.030000	239.170000	0.033160	0.000260	0.021440	0.019580	0.064330	0.119080	1.302000	...	0.169420	0.314820

8 rows × 23 columns

Figure 4.15: Details of parkinson data set

There are given some visualization of datasets that **shows relationship between different disease parameter and histogram for visualizing count of people.** This plotting shows why some types of person who are having a certain disease test result are more or less vulnerable to parkinson disease. It also indicates his or her future possibility of being parkinson disease patient. I used these plotting also to understand the classifier activities also. I used python **matplotlib** and **seaborn** library for these plotting purpose.



(a) Average vocal frequency range and count of non Parkinson and Parkinson people

(b) Maximum vocal frequency range and count of non Parkinson and Parkinson people

(c) Minimum vocal frequency range and count of non Parkinson and Parkinson people

(d) Several measures of variation in amplitude NHR,HNR and number of people having several ranges

(e) Maximum vocal fundamental frequency and number of people having several ranges

(f) Average vocal fundamental frequency and number of people having several ranges

Figure 4.16: Visualization of parkinson data set

4.3.2 K-Fold Cross Validation Results:

Table 4.3: Accuracy Results using Different ML Models

Number of Features	22
Logistic Regression Accuracy	84.10%
Support Vector Machine Accuracy	84.11%
Random Forest Accuracy	79.99%

4.3.3 Confusion Matrix:

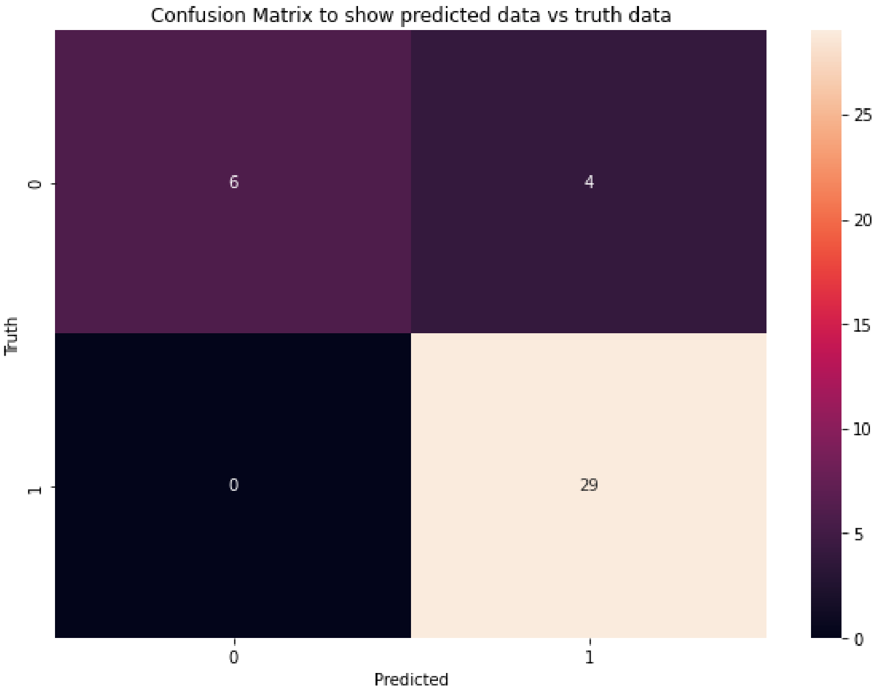


Figure 4.17: Confusion Matrix showing the contradiction between actual result and predicted result

4.3.4 Co-relation Matrix between features:

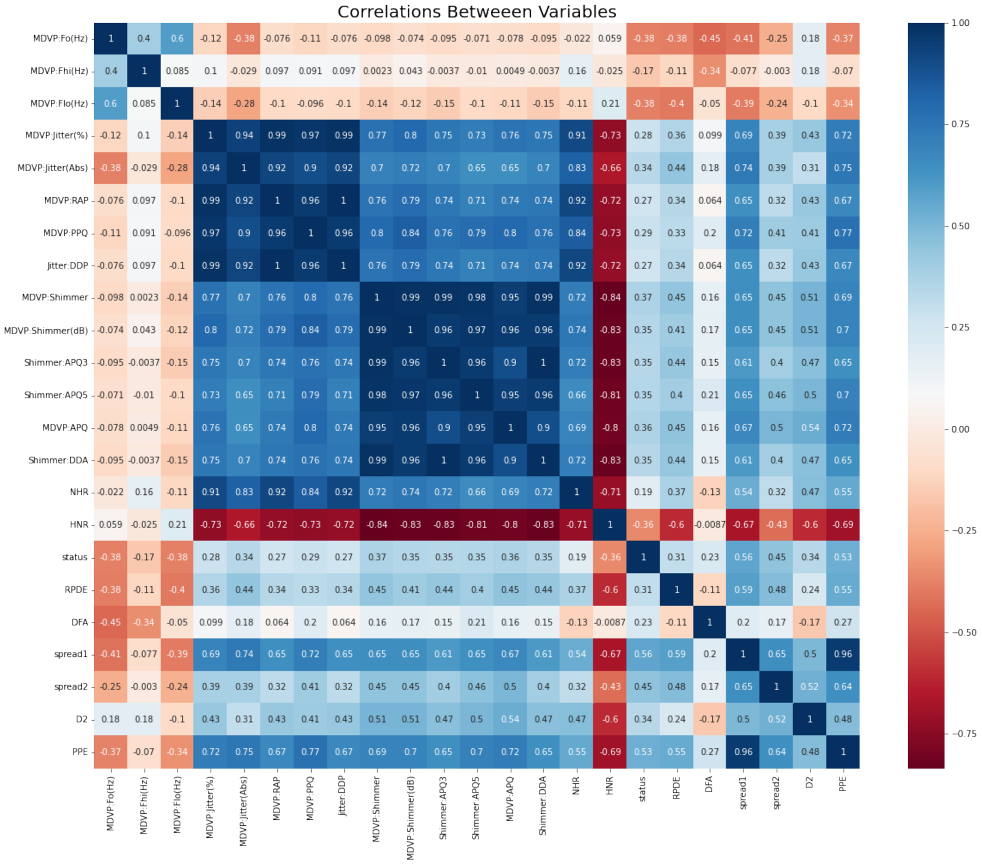
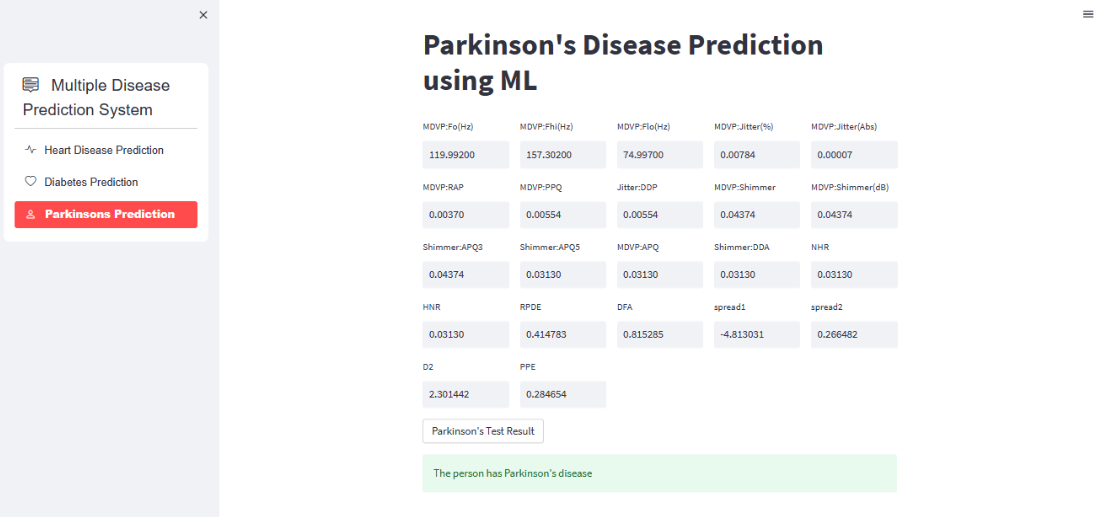


Figure 4.18: Shows the relation between all the features and their co-relationship

4.3.5 User Interface:

The below figure shows a Parkinson's disease holder patient and corresponding result. The input is taken from the user.



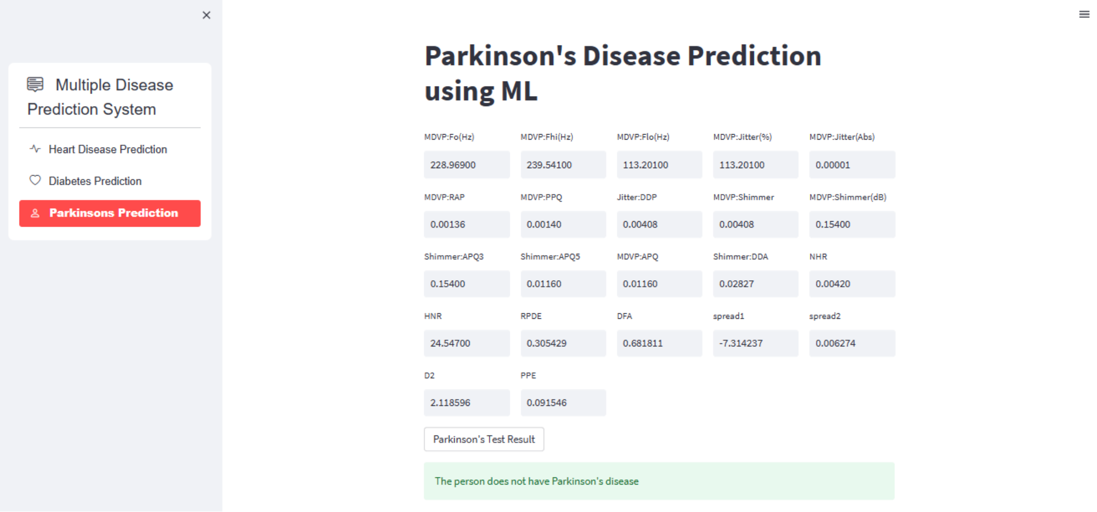
Parkinson's Disease Prediction using ML

MDVP:Fo(Hz): 119.99200
MDVP:Fhi(Hz): 157.30200
MDVP:Flo(Hz): 74.99700
MDVP:Jitter(%): 0.00784
MDVP:Jitter(Abs): 0.00007
MDVP:RAP: 0.00370
MDVP:PPQ: 0.00554
Jitter:DDP: 0.00554
MDVP:Shimmer: 0.04374
MDVP:Shimmer(dB): 0.04374
Shimmer:APQ3: 0.04374
Shimmer:APQ5: 0.03130
MDVP:APQ: 0.03130
Shimmer:DDA: 0.03130
NHR: 0.03130
HNR: 0.03130
RPDE: 0.414783
DFA: 0.815285
spread1: -4.813031
spread2: 0.266482
D2: 2.301442
PPE: 0.284654

The person has Parkinson's disease

Figure 4.19: User interface showing the prediction of parkinson patient

The below figure shows a person who doesn't have Parkinson's disease.



Parkinson's Disease Prediction using ML

MDVP:Fo(Hz): 228.96900
MDVP:Fhi(Hz): 239.54100
MDVP:Flo(Hz): 113.20100
MDVP:Jitter(%): 113.20100
MDVP:Jitter(Abs): 0.00001
MDVP:RAP: 0.00136
MDVP:PPQ: 0.00140
Jitter:DDP: 0.00408
MDVP:Shimmer: 0.00408
MDVP:Shimmer(dB): 0.15400
Shimmer:APQ3: 0.15400
Shimmer:APQ5: 0.01160
MDVP:APQ: 0.01160
Shimmer:DDA: 0.02827
NHR: 0.00420
HNR: 24.54700
RPDE: 0.305429
DFA: 0.681811
spread1: -7.314237
spread2: 0.006274
D2: 2.118596
PPE: 0.091546

The person does not have Parkinson's disease

Figure 4.20: User interface showing the prediction of non parkinson patient

5 Conclusion

My project work is about predicting multiple diseases using various classfication algorithm such as SVM(Support Vector Machine), Logistic Regression model and Random Forest. And these classifier algorithms provide a good accuracy in predicting diseases early. To conclude, the future purpose is to provide more disease prediction system in this one simple user interface.